On New Text Corpora For Minority Languages On The Helsinki korp.csc.fi Server

Rueter, Jack

2019-12-20

Rueter , J & Partanen , N 2019 , ' On New Text Corpora For Minority Languages On The
þÿ H e l s i n k i k o r p . c s c . f i S e r v e r ' , P a p e r p r e s e n t e d a t - ; 5 : B @ > = = 0 O ? 8 A L < 5
þÿ > A A 8 9 A : > 9 $ 5 4 5 @ 0 F 8 8 : > ? K B , ? @ > 1 ; 5 < K 8 ? 5 @ A ? 5 : B 8 2 K , U f a , R u s s
þÿ 2 7 / 1 1 / 2 0 1 9 - 2 9 / 1 1 / 2 0 1 9 p p . 3 2 3 6 .

http://hdl.handle.net/10138/309883

Дж. Рютер, Н. Партанен
г. Хельсинки, Финляндия

НОВЫЕ ТЕКСТОВЫЕ КОРПУСЫ МИНОРИТАРНЫХ ЯЗЫКОВ НА СЕРВЕРЕ KORP.CSC.FI.

Сервер Korp.csc.fi в Финляндии предоставляет текстовые корпусы нескольких разновидностей для многих языков – больших и малых. Инфраструктура Корп разработана шведской компанией Språkbanken в университете и Гетеборге, исходный код выпущен под лицензией MIT. Открытый характер системы позволяет легко переноситься в новую среду, и уже существуют многочисленные Корп установки. Тот, который мы обсуждаем, поддерживается языковым банком Финляндии.
Ключевые слова: корпус уральских языков меньшинств, минимальные данные для Korpa, корпус языков с ограниченными ресурсами.

J. M. Rueter, N. Partanen
Helsinki, Finland

ON NEW TEXT CORPORA FOR MINORITY LANGUAGES ON THE HELSINKI KORP.CSC.FI SERVER

The korp.csc.fi server in Finland provides text corpora of multiple varieties for numerous languages large and small. The Korp infrastructure is developed by the Swedish Språkbanken in the University and Gothenburg, and the source code is released under MIT license. Open nature of the systems makes it easily transferred into new environments, and there are already numerous Korp installations available. The one we discuss is maintained by the Language Bank of Finland.
Keywords: minority-language Uralic corpora, minimal data for Korp, low-resourced language corpora.

The korp.csc.fi server in Finland provides text corpora of multiple varieties for numerous languages large and small. The Korp infrastructure [Ahlberg et. al. 2013] is developed by the Swedish Språkbanken in the University and Gothenburg, and the source code is released under MIT license. Open nature of the systems makes it easily transferred into new environments, and there are already numerous Korp installations available. The one we discuss is maintained by the Language Bank of Finland.

In addition to the official languages of Finland, Finnish and Swedish, there is an “other languages” section featuring materials from Hurro-Urartian, Indo-European, Niger-Congo, Semitic, and Uralic language families. While the Indo-European and Niger-Congo materials are listed by language and focus on large modern languages (e.g. English, French, German, Russian, Spanish and Swahili), the Oracc collection features Hurro-Urartian, Semitic and other Mesopotamian materials, and the Uralic collections include materials from non-majority languages. It is the Uralic materials that will be described more in detail here.

The Uralic materials are divided according to licensed group: ERME, Fenno-Ugrica, SUS-kenttätyö, Wanca [Jauhiainen 2015], Kildin Saami, BeserCorp. The ERME, Fenno-Ugrica, Wanca and SUS-kenttätyö collections contain multilingual materials; Erme – mainly Erzya and Moksha literary materials, Fenno-Ugrica (National Library of Finland) represents pedagogical and literary materials from the 1920s through the early 1950s for ten Uralic languages, Wanca (Kone, FINCLARIN), in contrast, represents internet materials for 29 Uralic languages, and the SUS-kenttätyö (Finno-Ugrian Society Fieldwork) collection is just the beginning of Uralic fieldwork materials to be made available from the Pre-Soviet Era (normalized text search with fieldwork transcriptions and research-language translation, i.e. German, Russian, or other). The Kildin Saami sample corpus provides original unannotated orthographic text as are currently found in the Finno-Ugrica and Wanca collections. And finally, the BeserCorp provides a mixed encoding representation of Beserman Udmurt texts, where obvious Russian loanwords are written in Cyrillic script and other words are written in IPA-like Latin transcription with morphological segmentation and analyzed glossing.

Making materials available on a korp server, with concordance and more specific search options, presupposes some rudimentary language resources. First, there is the text, which may be any variety of orthographic, phonetic or other UTF-8 supported encoding. Second, the texts should be broken into sentence-like segments that eventually might be parsed syntactically, or annotated at other levels. Finally, each sentence is given a unique identification and tokenized with 1,2,3… unique identification of each individual token for each sentence (see figures 1–2). These are the minimal requirements, and any additional information including lemmas, part-of-speech, morpho-syntactic analysis, glosses and sentence translation are enhancement.

Figure 1. Illustrates the beginning “<sentence>” tag with attributes and values for unique identification as well as the original text and German translation. The attribute paragID for paragraph identifier along with the sent attribute are sufficient data for unique identification when accompanied by the text identifier. Superfluous data might be the pgNo (page number) and pgLi (page line) attributes, which align with reference provided in the printed texts, where the original transcriptions and translations are aligned at the line level.

```
<sentence paragID="1" sent="1" pgNo="0008" pgLi="1" orig_string="ruťš
šuę̌ turǐ̯li: «vetlan, voę̌, me ordę̌ gę̌ťitnǐ̯!»" deu="Der Fuchs sagt zum
Kranich: «Laß uns, mein Brüderchen, zu mir zu Besuch gehen!»">
```

Fig. 1.

Fig. 2. Provides an XML structure where the original-string tokens from fieldwork transcriptions have been copied to a line by line set of “<w/>” elements with “sID” string identification attributes.

```
<sentence paragID="1" sent="1" pgNo="0008" pgLi="1" orig_string="ruťš
šuę̌ turǐ̯li: «vetlan, voę̌, me ordę̌ gę̌ťitnǐ̯!»" deu="Der Fuchs sagt zum
Kranich: «Laß uns, mein Brüderchen, zu mir zu Besuch gehen!»">
<w word="руч" lemma="" pos="" msd="" sID="1" orig_string="ruťš"/>
<w word="шуö" lemma="" pos="" msd="" sID="2" orig_string="šuę̌"/>
<w word="турилы" lemma="" pos="" msd="" sID="3" orig_string="turǐ̯li"/
>
<w word=":" lemma=":" pos="CLB" msd="CLB" sID="4" orig_string=":"/>
<w word="«" lemma="«" pos="CLB" msd="CLB" sID="5" orig_string="«"/>
<w word="ветлан" lemma="" pos="" msd="" sID="6" orig_string="vetlan"/
>
<w word="," lemma="," pos="CLB" msd="CLB" sID="7" orig_string=","/>
<w word="воö" lemma="" pos="" msd="" sID="8" orig_string="voę̌"/>
<w word="," lemma="," pos="CLB" msd="CLB" sID="9" orig_string=","/>
<w word="ме" lemma="" pos="" msd="" sID="10" orig_string="me"/>
<w word="ордö" lemma="" pos="" msd="" sID="11" orig_string="ordę̌"/>
<w word="гöсьтиӟны" lemma="" pos="" msd="" sID="12"
orig_string="gę̌ťitnǐ̯"/>
<w word="!" lemma="!" pos="CLB" msd="CLB" sID="13" orig_string="!"/>
<w word="»" lemma="»" pos="CLB" msd="CLB" sID="14" orig_string="»"/>
</sentence>
```

Fig. 2.

Each token has then been copied to both a “word” and “orig_string” attribute, where Komi-Zyrian Cyrillic normalization has been applied to the “word” attribute value and the fieldwork transcription value has been retained in the orig_string attribute. Initial work has been carried out for the Mordvin languages Erzya and Moksha by these authors using simple lines of perl script for ordered normalization, but continued work will move towards practices used in early Modern English Normalization [Hämäläinen 2019].

Figure 3. illustrates the VRT representation of this minimal tokenization and string identifier data. The first line of the individual columns, in red, is merely an indicator of the data type. Only punctuation symbols have been lemmatized with part-of-speech tagging and morphological analyses. The absence of data value is indicated by a stroke below symbol “_”.

Wordform	Lemma	PoS	Msd	StringId	Transcription
руч	_	_	_	1	ruťš
шуö	_	_	_	2	šuę̌
турилы	_	_	_	3	turǐ̯li
:	:	CLB	CLB	4	:
«	«	CLB	CLB	5	«
ветлан	_	_	_	6	vetlan
,	,	CLB	CLB	7	,

Fig. 3.

If the language has a little more resources, which is the case for Komi-Zyrian, lemmas, parts-of-speech and morphological analyses may also be had with minimal manual work (see Fig. 4.). Even with current language technology that is available, however, this manual correction does not scale well into larger corpora, for which the further development of the available tools seems to be the best approach. That said, even a small manually annotated corpora can be very valuable in many tasks, and such work is always very recommendable. Pipelines described in connection to different treebanks [Partanen 2018, Rueter 2018], which combine finite state transducers into manual disambiguation, can be easily adapted in this context as well.

Wordform	Lemma	PoS	Msd	StringId	Transcription
руч	руч	NOUN	N.Sg.Nom	1	rutš̆
шуö	шуны	VERB	V.Ind.Prs.Sg3	2	šuę
турилы	тури	NOUN	N.Sg.Dat	3	turi̯li
:	:	PUNCT	CLB	4	:
«	«	PUNCT	CLB	5	«
ветлан	ветлыны	VERB	V.Ind.Fut.Sg2	6	vetlan
,	,	PUNCT	CLB	7	,

Fig. 4.

As in Fig. 3. before it, the first row of the columns in Fig. 4. is merely intended to illustrate the data type. The order of the columns, it should be noted, is documented in each project separately, i.e. there is no specific ordering to the individual columns. Hence a similar project from Moksha might be illustrated with column ordering the same as in the Universal Dependencies forth-coming release [15 Nov. 2019; cf. Nivre, et al, 2019] (see Fig. 5.). Here the ordering applied is by column as follows: 1= string id, 2 = word form or token, 3 = lemma, 4 = part-of-speech, 5 = alternate encoding, 6 = features, 7 = dependency, 8 = dependency relation, 9 = NA, and 10 = miscellaneous.

```
<sentence sent_id="MishaninaValentina_LiendenyOchkonyasa_Moksha-1972-No2-pp38-39:36" text="Лётчикне, улема, кядьса токсесазь коволнятнень.">
1   Лётчикне    лётчик  NOUN    N   Case=Nom|Definite=Def|Number=Plur    6   nsubj   _   SpaceAfter=No
2   ,           ,       PUNCT   CLB _                                   3   punct   _   _
3   улема       улема   PART    Pcle    _                               6   advmod  _   SpaceAfter=No
4   ,           ,       PUNCT   CLB _                                   3   punct   _   _
5   кядьса      кядь    NOUN    N   Case=Ine|Definite=Ind|Number=Plur,Sing  6   obl     _   _
6   токсесазь   токсемс VERB    V   Mood=Ind|Number[obj]=Plur|Number[subj]=Plur|Person[obj]=3|Person[subj]=3|Tense=Pres|Valency=2  0   root    _   _
7   коволнятнень    коволня NOUN    N   Case=Gen|Definite=Def|Number=Plur    6   obj
```

Fig. 5.

This kind of additional information can be added at the collection, text, paragraph, sentence or token level, and all of these attributes can be queried in the search interface.

Korp uses VRT XML format, in which each tokenized sentence is represented in a tabular structure within XML's text element. This tabular structure is split so that each line represents a wordform, and the annotations about that wordform are gathered at different columns at the same row (cf. fig 3–4.). In principle the number of columns is arbitrary, and any information could be added this way.

The fields and tags in Korp can be translated within a configuration file, and this provides many possibilities for a multilingual interface. At the moment the whole interface is available in Finnish, Swedish and English. There are three search types: Simple, Extended and Advanced. The Advanced search mode uses CQP, which is a widely used and very expressive query language. Besides search form itself, Korp also has an API, which allows all the same search actions and the ordinary interface. Korp API can also be accessed through a Python package [Hämäläinen 2018].

It is relatively easy to add new languages and corpora to this infrastructure, especially in cases where the licenses are easily defined and the ownership of data is clear. It is also possible to add materials to the infrastructure that need various restrictions. One alternative is to make the data available for academic

research, which means that anyone with an account in a university that belongs to the HAKA-system can use the materials after login. This is relatively strong authentication, with a problem that the universities within this system are mainly located in Europe. However, for partially controlled use it is currently the best option offered in this infrastructure. It is also possible to make the materials available on request, which would demand an application and explanation of intended use. Although laborious to apply and process, this access method would in principle be open to anyone who fills the conditions corpus creators have set for their materials.

Table 1. provides a statistical summary for token and phrase coverage of language corpora on the Helsinki korp.csc.fi server.

English name	Project	Number of word forms	Number of sentences
Eastern & Meadow Mari	Fenno-Ugrica, Wanca	2,968,864	272630
Erzya	Erme, Fenno-Ugrica, SUS-Kenttätyö, Wanca	2,042,388	201,104
Moksha	Erme, Fenno-Ugrica, SUS-Kenttätyö, Wanca	1,673,480	142,248
Tundra Nenets	Fenno-Ugrica,Wanca	1,570,794	226,736
Hill or Western Mari	Fenno-Ugrica, Wanca	1,267,860	120,925
Mansi	Fenno-Ugrica, Wanca	936,869	120,598
Võro	Wanca	903,900	73,043
Veps	Fenno-Ugrica, Wanca	655,549	91,721
Khanty	Fenno-Ugrica, Wanca	620,815	72,934
Udmurt	Wanca	613,043	57,873
Ingrian	Fenno-Ugrica, Wanca	465,672	65,823
Selkup	Fenno-Ugrica	386,354	76,335
Komi-Zyrian	SUS-Kenttätyö, Wanca	233,596	22,286
Olonets-Karelian	Wanca	128,916	10,442
Komi-Permyak	Wanca	93,010	8,595
Skolt Sami	Wanca	91,154	8,402
Udmurt Beserman	BeserCorp	47,229	5,234
Karelian	Wanca	33,609	2,849
Ludic	Wanca	8,802	802
Kildin Sami	Kildin Saami Sample, Wanca	2,515	222
Nganasan	Wanca	2,161	67
Votic	Wanca	191	22

Table 1. Coverage for minority Uralic languages spoken in Russia

The infrastructure described in this article offers a flexible and robust platform for various types of linguistic corpora. Although numerous alternatives exist for tools described here, many with much larger resources, the relationship of different corpus providers should probably be best seen as complementing than competing. For example, for Komi-Zyrian komicorpora-website maintained by FU-Lab (FU-Lab 2019), is certainly the source of widest and first priority. Similarly, for Sami languages the materials provided by Giellatekno are in many ways crucial. However, wider representation of partially overlapping resources in differences interfaces, with possibly differing annotation schemes, is perhaps ultimately only beneficial for the field at large.

Looking onward, there are several questions for which the best practices still have to be adopted, and possibly invented. The versioning of the corpora, which would allow easy and flexible updating, is one of the tasks where further development is certainly needed. This, however, is not so much a technical problem, but relates to the practices and responsibility sharing around the corpus maintenance. As far as we see, consistent versioning, clear licensing, persistent identifiers and automatic validation must somehow be integrated to the workflows how these corpora are maintained and updated in the future.

References

Ahlberg Malin, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, and Jonatan Uppström. // Korp and Karpa bestiary of language resources: the research infrastructure of Språkbanken. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), P. 429–433. 2013.

FU-Lab. Корпус коми языка: [Электронный ресурс]. URL: http://komicorpora.ru/ (дата обращения: 28.10.2019).

Hämäläinen M., Säily T., Rueter J., Tiedemann J., & Mäkelä E. Revisiting NMT for normalization of early English letters. In B. Alex, S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.),

Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. P. 71–75. (ACL Anthology; No. W19-25). Stroudsburg: Association for Computational Linguistics, 2019.

Jauhiainen, Tommi, Heidi Jauhiainen, and Krister Lindén // The Finno-Ugric languages and the internet project. First International Workshop on Computational Linguistics for Uralic Languages Proceedings of the Workshop. Septentrio Academic Publishing. 2015.

Mika Hämäläinen. (2018, January 9). Python Korp Library (Version v1). Zenodo. http://doi.org/10.5281/zenodo.1143374

Nivre, Joakim; Abrams, Mitchell; Agić, Željko; et al., 2019, *Universal Dependencies 2.4*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2988.

Partanen, Niko, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. The First Komi-Zyrian Universal Dependencies Treebanks // Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium, P. 126–132.

Rueter, Jack M. & Tyers, Francis. Towards an open-source universal-dependency treebank for Erzya. Proceedings of the International Workshop for Computational Linguistics of Uralic Languages. Helsinki, Finland, 08/01/2018–09/01/2018.